Issues in Optical Diffraction Theory

Klaus D. Mielenz

National Institute of Standards and Technology (ret.)
Gaithersburg, MD 20899-8440

kmielenz@wildblue.net

This paper focuses on unresolved or poorly documented issues pertaining to Fresnel's scalar diffraction theory and its modifications. In Sec. 2 it is pointed out that all thermal sources used in practice are finite in size and errors can result from insufficient coherence of the optical field. A quarter-wave criterion is applied to show how such errors can be avoided by placing the source at a large distance from the aperture plane, and it is found that in many cases it may be necessary to use collimated light as on the source side of a Fraunhofer experiment. If these precautions are not taken the theory of partial coherence may have to be used for the computations.

In Sec. 3 it is recalled that for near-zone computations the Kirchhoff or Rayleigh-Sommerfeld integrals are applicable, but fail to correctly describe the energy flux across the aperture plane because they are not continuously differentiable with respect to the assumed geometrical field on the source side. This is remedied by formulating an improved theory in which the field on either side of a semi-reflecting screen is expressed as the superposition of mutually incoherent components which propagate in the opposite directions of the incident and reflected light. These components are defined as linear combinations of the Rayleigh-Sommerfeld integrals, so that they are rigorous solutions of the wave equation as well as continuously differentiable in the aperture plane. Algorithms for using the new theory for computing the diffraction patterns of circular apertures and slits at arbitrary

distances z from either side of the aperture (down to $z = \pm 0.0003 \lambda$) are presented, and numerical examples of the results are given. These results show that the incident geometrical field is modulated by diffraction before it reaches the aperture plane while the reflected field is spilled into the dark space. At distances from the aperture which are large compared to the wavelength λ these field expressions are reduced to the usual ones specified by Fresnel's theory. In the specific case of a diffracting half plane the numerical results obtained were practically the same as those given by Sommerfeld's rigorous theory.

The modified theory developed in this paper is based on the explicit assumption that the scalar theory of light cannot explain polarization effects. This premise is justified in Sec. 4, where it is shown that previous attempts to do so have produced dubious results.

Key words: bidirectional scalar diffraction; continuously differentiable field components; Fresnel; Kirchhoff; near-field; partial coherence; plane apertures; polarization; pseudo-vectorial theories; Rayleigh; Sommerfeld; transmission coefficients.

Accepted: January 27, 2009

Available online: http://www.nist.gov/jres

1. Introduction

Calculations pertaining to the diffraction of light by an aperture are typically based on geometric assumptions as illustrated in Fig. 1. A plane screen $\mathscr{D}$ containing an aperture $\mathscr{A}$ of width $2w$ is illuminated by a quasi-monochromatic source $\mathscr{S}$ of width $2s$ and circular wave number $k = 2\pi / \lambda$, P_0 is a source point, Q is a point inside the aperture, and P is the point at which diffraction is observed. A point inside the aperture (usually the center) is chosen as the origin O, and cartesian or polar coordinates are used so that

$$P_0 = (x_0, y_0, z_0) = r_0(\cos\phi_0\sin\theta_0, \sin\phi_0\sin\theta_0, \cos\theta_0),$$

$$Q = (\xi, \eta, 0) = q(\cos\chi, \sin\chi, 0), \qquad (1)$$

$$P = (x, y, z) = r(\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta).$$
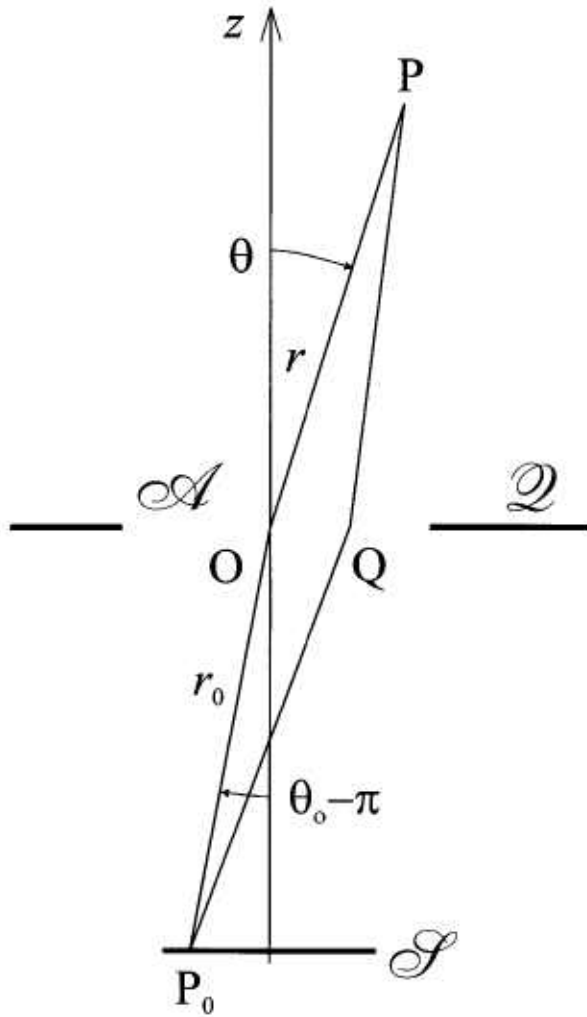

Fig. 1. Geometry pertaining to Fresnel's diffraction integral.

The first mathematical theory of diffraction was derived in Fresnel's 1819 *Memoir on the Diffraction of Light* [1] from a surprisingly minimal set of assumptions:

Monochromatic light is a harmonic wave motion that can be described in terms of a scalar wave function, $U(P)\exp(-i\omega t)$, where $|U(P)|^2$ is the irradiance at P, $\omega = 1/kc$, is the circular frequency $k = 2\pi/\lambda$, is the circular wave number and c is the speed of light.

The summation of Huygens' wavelets is best carried out on a wave front that coincides with the aperture; otherwise the computations will be too complicated.

A plane aperture is assumed, which implies that the source must be distant as well as paraxial $(r_0 > 2w^2/\lambda$, $\cos\theta_0 \sim 0)$,[1] so that the incident wave front can, within acceptable tolerances, be assumed to lie inside the aperture.

The field on the source side of the screen is not affected by the presence of the screen (St. Venant's hypothesis).[2] For an infinitesimally small source with radiant intensity I which is located at the point P_0 in Fig. 1, this leads to $U(Q) = \sqrt{I}\exp(ikP_0Q)/P_0Q$.

The Huygens' wavelets originating at Q are anisotropic spherical waves that oscillate a quarter period ahead of the incident field and whose amplitudes are scaled by the factor $1/\lambda$. As the corresponding effect at the point of observation P is attributable to the central Fresnel zone at Q acting alone, there is no need to know the nature of this anisotropy and hence it follows that $dU(P) = (i/\lambda)U(Q)\exp(ikQP)/QP$.

On account of the denominator λQP in the last expression it is necessary to assume that QP is a large distance $(r \gg \lambda)$, and that P must also be paraxial $(\cos\theta \approx 1)$ as the direction of observation cannot be substantially different from the direction of incidence.

Combining these expression, Fresnel found

$$U_F(P) = \frac{i\sqrt{I}}{\lambda}\int_{\mathscr{A}} dQ\,\frac{\exp[ik(P_0Q + QP)]}{P_0Q\,QP}$$

$$\sim \frac{i\sqrt{E_{geom}}}{\lambda r}\exp[ik(r_0 + r)]\int_{\mathscr{A}} dQ\exp[ik\Delta(Q)],\quad z \gg \lambda, \qquad (2)$$

where $E_{geom} = I/r_0^2$ is the incident geometrical irradiance of the aperture plane and $\Delta(Q)$ is the path difference $(P_0Q + QP) - (r_0 + r)$.

Although a relic of the early 19th century, the Fresnel diffraction integral (2) is still used today in its original

---

[1] This inequality for r_0 follows from Eq. (4a) for $s = 0$.

[2] Jean Claude Saint-Venant (1797-1866) taught mathematics at the École des Ponts et Chaussées in Paris, where Fresnel studied engineering between 1806 and 1809.

form and has remained a most useful, reliable tool for diffraction calculations that have consistently yielded results which agree with experience and are well documented [2]. Accordingly, it is commonly regarded as a cornerstone of diffraction theory and its validity for mid-field applications remains unchallenged. On the other hand, the classical theory of optical diffraction is limited in scope as well as physical significance by several issues which are still unresolved, misunderstood, or poorly documented. For example:

The assumption of an infinitesimally small isotropic point source is seldom justified. All thermal sources are finite in size, so that errors may arise from an insufficient coherence of the optical field. As shown in Sec. 2, below, this can be avoided by invoking the theory of partial coherence, or by negating the error in the first place by designing a diffraction experiment so that the aperture illumination is "almost" coherent.

As Fresnel's integral is not applicable for computations in the near zone, it would seem that in this case the more accurate Rayleigh-Sommerfeld or Kirchhoff boundary-value integrals [Eqs. (5a-c), below] can be used with confidence. This is, however, not the case as these integrals do not correctly describe the field in the proximity of the aperture screen. In Sec. 3, this problem is solved by constructing an improved theory in which St. Venant's hypothesis is abandoned and the field on both sides of the screen is expressed in terms of linear combinations of the Rayleigh-Sommerfeld integrals.

The standard solutions of the Fresnel diffraction integral (2) for circular apertures and apertures bounded by straight edges are models of mathematical elegance, but notoriously difficult to evaluate in practice. The numerical methods used in this paper not as elegant, but easier to use. As they are rigorously correct, they can be used at arbitrary distances from the screen, and this is recommended.

Initially, Fresnel believed that light is akin to longitudinal sound waves but prior to the completion of his theory he discovered, in collaboration with Arago,[3] that it is a transverse wave. He pondered the obvious question how this discovery affected the theory of diffraction and concluded that "*the arguments and*

computations contained in the Memoir harmonize quite as well with this new hypothesis as with the preceding, because they are independent of the actual directions of the vibrations." In other words, the scalar approach makes it impossible to describe the diffraction of polarized light. This is confirmed in Sec. 4, below.

2. Coherence Issues

Extended thermal sources employed in diffraction experiments are used either by themselves or followed by a limiting aperture. In either case, the diffraction pattern due to the principal aperture $\mathscr{A}$ can be evaluated using a generalized Fresnel integral derived by this author from the general equations for the propagation of cross-spectral density in a partially coherent optical field [3]. The case of an incoherent source used by itself is a straightforward generalization of Eq. (2) and leads to the following expression for the irradiance at the point P in Fig. 1,

$$E(P)=\frac{1}{(\lambda r)^2}\int_{\mathscr{S}} dP_0\, L(P_0)\left|\int_{\mathscr{A}} dQ\exp[ik(P_0Q+QP)]\right|^2, \text{(3a)}$$

where $L(P_0)$ is the source radiance. In the case of a thermal source followed by an aperture located in the plane $\mathscr{S}$, it is necessary to use the van Cittert-Zernike theorem to calculate the cross-spectral density $W(P_0, P_0', \omega)$ on $\mathscr{S}$ and then find E(P) from the expression

$$E(P) = \frac{1}{\lambda^4 r^2}\int_{\mathscr{S}} dP_0 dP_0'\, \frac{W(P_0, P_0', \omega)}{P_0Q\,P_0Q'}$$
$$\int_{\mathscr{A}} dQdQ'\, \exp[ik(P_0Q+QP-P_0'Q'+Q'P], \text{(3b)}$$

where P_0 and P'_0 are two points on $\mathscr{S}$, Q and Q' are points in $\mathscr{A}$, and ω is the circular frequency of the light. These equations were applied in Ref. [4] to the specific case of concentric circular apertures and yielded complicated but closed solutions in terms of Lommel functions.

The use of Eqs. (3a) and (3b) for practical computations is a very tedious task. Therefore, it may be desirable to avoid the need for these computations in the first place by designing diffraction experiments so that the aperture illumination will be "almost" coherent and the Fresnel integral (2) can still be used, in spite of the finite size of a given source. This can be done as follows, assuming the worst case of a totally incoherent source as implied by Eq. (3a), considering a concentric arrangement of source and aperture as in Fig. 1, and then applying the quarter-wave criterion so that the path lengths P_0Q do not vary by more than $\lambda/4$.

---

[3] At the time of this collaboration with Fresnel, Francois Jean Dominique Arago (1786-1853) was Secretary of the Paris Observatory. For a brief time in the Spring of 1848 he served as the Head of the Second French Republic.

According to Fig. 1, the extreme values of P_0Q are

r_0 and $\sqrt{r_0^2 + (s+w)^2} \approx r_0 + (s+w)^2/2r_0$, and thus the desired criterion is $(s+w)^2/r_0 \le \lambda/2$ or

$$r_0 \ge \frac{2(s+w)^2}{\lambda}. \tag{4a}$$

As shown in Table 1, the corresponding minimum distances $r_{0,\min}$ vary significantly with the sizes of source and aperture, ranging from a few millimeters for small sources and apertures to hundreds of meters for large ones. In the latter case Eq. (4a) can only be satisfied by placing the source in the focal plane of a collimator lens, as on the source side of a Fraunhofer diffraction experiment.[4]

Table 1. Smallest permissible source-aperture distance $r_{0,\min}$ defined by Eq. (3a) for selected aperture sizes $2w$ (left column) and source sizes $2s$ (top row), $\lambda = 1$ μm

$2w/2s$	0.01 mm	0.1 mm	1 mm	10 mm
0.1 mm	6 mm	20 mm	605 mm	50.1 m
1 mm	510 mm	605 mm	2 m	60.5 m
10 mm	50 m	51 m	60.5 m	200 m

The coherence criterion (4a) assures a coherent and uniform aperture field so that the path lengths ($P_0Q + QP$) are all effectively equal to ($r_0 + QP$), as if the incident field is a plane wave propagating in the direction of the z-axis and $U_{\text{geom}} = \sqrt{E_{\text{geom}}}\exp(i\,kz)$. Under these conditions Eq. (3a) is reduced to a product of independent integrals over source and aperture, and it is easy to show that the first of these, $\int_{\mathscr{S}} dP_0\, L(P_0)$ is equal to the geometrical irradiance of the aperture plane if Lambert's law is assumed. That is,

$$E(P) = \frac{1}{(\lambda r)^2} \int_{\mathscr{S}} dP_0\, L(P_0) \times \left| \int_{\mathscr{A}} dQ \exp[i\,k(r_0 + QP)] \right|$$

$$= \frac{E_{\text{geom}}}{(\lambda r)^2} \left| \int_{\mathscr{A}} dQ \exp(ikQP) \right|^2, \tag{4b}$$

which is the squared modulus of Eq. (2) for $P_0Q = r_0$.

---

[4] Equation (4a) is more stringent than a well-known coherence criterion attributed to Verdet (1869) which specifies different minimum distances, $r_{0,\min} = 8sw/\lambda$. These are also quite large for large apertures and sources.

3. Rigorous Theory

3.1 Background

Fresnel was aware that the spherical wavelets assumed in his derivation of Eq. (2) cannot be iso-tropic, because otherwise light would also travel back toward the source. To avoid this contradiction, so-called inclination factors can be introduced to assure that the amplitudes of the wavelets are zero in the reverse direction. As mentioned earlier, Fresnel did not know the form of these factors and simply omitted them, assuming correctly that they are not needed in a paraxial theory.

The question of inclination factors was answered by the publication of Helmholtz' theorem in 1859 [5].[5] Whereas Fresnel had only stipulated that the screen $\mathscr{D}$ in Fig. 1 must be large enough to prevent the leakage of light around its edges, Helmholtz imagined it to be an infinitely large, closed surface which does not contain the primary source, and then invoked Green's formula to express the diffracted field $U(P)$ as a surface integral of the form prescribed by Fresnel's first assumption. This provided the missing inclination factors in the form of the normal derivatives $\partial U(Q)/\partial n$, which vanish when the light propagates in the direction tangential to the screen. In short, Helmholtz' theorem states that the diffracted field is confined to the inside of the surface $\mathscr{S}$ but null on the outside, and thus it merges Huygens' principle and St. Venant's hypothesis into one.

The principal solutions of Helmholtz' theorem are the familiar Rayleigh-Sommerfeld and Kirchhoff diffraction integrals [6,7,8],

$$U_{\text{RS}}^{(p)}(P) = -\frac{1}{2\pi} \int_{\mathscr{A}} dQ \frac{\partial U_{\text{geom}}(Q)}{\partial z} \frac{\exp(ikQP)}{QP}$$

$$= -\frac{ik\sqrt{E_{\text{geom}}}}{2\pi} \int_{\mathscr{A}} dQ \frac{\exp(ikQP)}{QP}, \quad z \ge 0, \tag{5a}$$

$$U_{\text{RS}}^{(s)}(P) = \frac{1}{2\pi} \int_{\mathscr{A}} dQ\, U_{\text{geom}}(Q) \frac{\partial}{\partial z}\left(\frac{\exp(ikQP)}{QP}\right)$$

$$= \frac{1}{ik}\frac{\partial U_{\text{RS}}^{(p)}}{\partial z}, \quad z \ge 0, \tag{5b}$$

$$U_{\text{K}}(P) = \frac{1}{2}[U_{\text{RS}}^{(p)}(P) + U_{\text{RS}}^{(s)}(P)], \tag{5c}$$

where the right-most expressions in (5a,b) were obtained by assuming that the coherence criterion (4a)

---

[5] Helmholtz' theorem was formulated as a mathematical theory of organ pipes, and thus also implies the assumption of longitudinal waves.

is satisfied so that $\sqrt{E_{\text{geom}}}\exp(ikz)$ could be substituted for $U_{\text{geom}}(Q)$. It is well known and easy to show that in the mid zone these expressions are all reduced to the Fresnel integral (2) so that

$$U_{RS}^{(p)}(P) = U_{RS}^{(s)}(P) = U_K(P) = U_F(P), \quad z \gg \lambda. \quad (5d)$$

It is equally well known that the Rayleigh-Sommerfeld integrals reproduce the boundary values assumed in their derivation,

$$\frac{\partial U_{RS}^{(p)}(Q)}{\partial z} \equiv ik\sqrt{E_{\text{geom}}},$$

$$U_{RS}^{(s)}(Q) \equiv \sqrt{E_{\text{geom}}}, \quad Q \in \mathscr{A}, \quad (5e)$$

but not the corresponding values, $U_{RS}^{(p)}(Q) = \sqrt{E_{\text{geom}}}$ and $\partial U_{RS}^{(s)}(Q)/\partial z = ik\sqrt{E_{\text{geom}}}$. Therefore, the wave functions defined by Eqs. (5a-c) are different from one another in the near zone and not continuously differentiable in the aperture, so that none expresses the diffracted field as an analytical continuation of the assumed geometrical field on the source side.

In order to overcome this failure of the classical boundary-value theories to describe a smooth flow of energy across the aperture plane, so-called "rigorous" theories have been formulated in which St. Venant's hypothesis is abandoned and it is assumed that the incident geometrical is modified by diffraction before it reaches the screen. Hence, its values on the source side can be determined by postulating that the overall field is continuously differentiable inside the aperture. The most important, and by far most successful treatment of this type is Sommerfeld's rigorous theory of diffraction by a half plane [7, 9], which is expressed in closed form and involves no approximations of any sort. Other examples are the Rayleigh-Bouwkamp [6, 10] and Levine-Schwinger [11] theories, which were intended to define the transmission coefficients of very small apertures but are approximations based on assumed aperture field distributions expressed as a series of algebraic functions. Undoubtedly, the rigorous treatment of diffraction problems is more powerful than the boundary-value approach but unfortunately some of the most common problems, such as diffraction by circular apertures or slits of arbitrary sizes, have so far not been solved rigorously.

The reasons for this deficiency of the prior literature were analyzed in a recent study [12] which included a numerical comparison of the respective half-plane results obtained from the Sommerfeld and Rayleigh-Sommerfeld theories. The theory derived in the following subsection is based on the findings of this study, and it is important that it involves concepts not found in other theories:

The frequently encountered association of the Rayleigh-Sommerfeld integrals with parallel or perpendicularly polarized light is abandoned and nothing is assumed about them, except that they are independent solutions of the wave equation which satisfy the respective boundary conditions specified in Eq. (5e), above.

Similarly, the customary distinction between "black" and "metallic" screens is abandoned. The physical nature of the screen is not specified, except that it has a certain reflectance $|\rho|^2$ which can be used as a scale factor to describe different types of screens.

In the general case of a partially reflecting screen there will be two mutually incoherent diffraction patterns, one in the forward direction of the incident light and the other in the reverse direction of the reflected light. The latter is usually not observable, but must be taken into account because its presence affects the radiant flux transmitted in the forward direction.

3.2 Derivations

Consider the diffraction of light by a plane aperture $\mathscr{A}$ in an infinite, infinitesimally thin screen $\mathscr{D}$ which is illuminated by a source $\mathscr{S}$, as shown in Fig. 1. Assume that $\mathscr{D}$ is a specular reflector with amplitude reflectance ρ and that the coherence condition (4a) is satisfied, so that the incident geometric field and its reflection can be expressed in the form of plane waves,

$$U_{\text{geom}} = \sqrt{E_{\text{geom}}}\exp(ikz) \text{ and } \hat{U}_{\text{geom}} = \rho\sqrt{E_{\text{geom}}}\exp(-ikz).$$

Under these conditions, the diffracted field will likewise be composed of mutually incoherent components $U(P)$ and $\hat{U}(P)$, which travel in opposite directions. The task at hand is to define these components so that each is continuously differentiable inside the aperture.

Using a normalized notation so that $U(P) = \sqrt{E_{\text{geom}}}\,u(P)$ and $\hat{U}(P) = \sqrt{E_{\text{geom}}}\,\hat{u}(P)$, we express $u(P)$ and $\hat{u}(P)$ in the form of the trial solutions,

$$u(P) = \begin{cases} u_+(P), & z > 0 \\ \exp(ikz) + u_-(P), & z < 0 \end{cases}$$

$$\hat{u}(P) = \rho \begin{cases} \hat{u}_+(P), & z > 0 \\ \hat{u}_-(P), & z < 0 \end{cases} \qquad (6a)$$

where $u_\pm(P)$ and $\hat{u}_\pm(P)$ are initially unknown field components which are attributed to diffraction effects.

As these unknown quantities must have certain prerequisite properties such as obeying the wave equation as well as the infinity and edge conditions [7, 13], and because the Rayleigh-Sommerfeld integrals (5a, b) have these properties, $u_\pm(P)$ and $\hat{u}_\pm(P)$ can be defined as linear combinations of the form

$$u_\pm(P) = a\, u_{RS}^{(p)}(P) \pm b\, u_{RS}^{(s)}(P),$$

$$\hat{u}_\pm(P) = \hat{a}\, u_{RS}^{(p)}(\hat{P}) \pm \hat{b}\, u_{RS}^{(s)}(\hat{P}), \qquad (6b)$$

where $\hat{P} = (x, y, -z)$ is the reflection of the point of observation P, defined so that $u_{RS}^{(p,s)}(\hat{P})$ is a valid expression when $z < 0$.

The coefficients a, b and $\hat{a}$, $\hat{b}$ in (6b) can now be determined by postulating that the forward field is equal to Fresnel's integral $u_F(P)$ in the positive mid zone and equal to the incident geometrical field in the negative mid zone. The respective values for the reverse field are assumed to be zero and $U_F(\hat{P})$. That is,

$$u(P) = \begin{cases} (a+b)u_F(P) = u_F(P), & z \gg \lambda \\ \exp(ikz) + (a-b)u_F(P) = \exp(ikz), & -z \gg \lambda \end{cases} \qquad (6c)$$

$$\hat{u}(P) = \rho \begin{cases} (\hat{a}+\hat{b})u_F(\hat{P}) = 0, & z \gg \lambda \\ (\hat{a}-\hat{b})u_F(\hat{P}) = u_F(\hat{P}), & -z \gg \lambda \end{cases} \qquad (6d)$$

where Eq. (5d) was used to let $u_{RS}^{(p)} = u_{RS}^{(s)} = u_F$ when z/λ is large.

Equations (6c,d) are satisfied if $a = b = \frac{1}{2}$, $\hat{a} = -\hat{b} = \frac{1}{2}$, so that

$$u(P) = \begin{cases} u_K(P), & z > 0 \\ \exp(ikz) + \hat{u}_K(\hat{P}), & z < 0 \end{cases}$$

$$\hat{u}(P) = \rho \begin{cases} \hat{u}_K(P), & z > 0 \\ u_K(\hat{P}), & z < 0 \end{cases} \qquad (6e)$$

where

$$u_K(P) = \frac{1}{2}[u_{RS}^{(p)}(P) + u_{RS}^{(s)}(P)],$$

$$\hat{u}_K(P) = \frac{1}{2}[u_{RS}^{(p)}(P) - u_{RS}^{(s)}(P)]. \qquad (6f)$$

It remains to show that these expressions are continuously differentiable inside the aperture. In the case of $u(P)$, this will be the case if

$$\frac{1}{2}[u_{RS}^{(p)}(Q) + u_{RS}^{(p)}(Q)] =$$
$$1 + \frac{1}{2}[u_{RS}^{(p)}(Q) - u_{RS}^{(p)}(Q)], \qquad (6g)$$

$$\frac{1}{2}\frac{\partial}{\partial z}[u_{RS}^{(p)}(Q) + u_{RS}^{(p)}(Q)] =$$
$$ik + \frac{1}{2}\frac{\partial}{\partial(-z)}[u_{RS}^{(p)}(Q) - u_{RS}^{(p)}(Q)], \qquad (6h)$$

which according to Eq. (5e) is true. This argument can be repeated to prove that, likewise, $\hat{u}(P)$ is continuously differentiable when P = Q.

Equations (6e) represent the main result of this Section. The corresponding forward and reverse irradiances of the field are

$$E(P)/E_{geom} = |u(P)|^2 = \begin{cases} |u_K(P)|^2, & z > 0 \\ |\exp(ikz) + \hat{u}_K(\hat{P})|^2, & z < 0 \end{cases} \qquad (6i)$$

$$\hat{E}(P)/E_{geom} = |\hat{u}(P)|^2 = \begin{cases} |\hat{u}_K(P)|^2, & z > 0 \\ |u_K(\hat{P})|^2, & z < 0 \end{cases} \qquad (6j)$$

Like $u(P)$ and $\hat{u}(P)$ themselves, these expressions are continuously differentiable inside the aperture and thus imply a smooth, bidirectional flow of energy from one side to the other. It should be noted that, in these expressions, the roles of Kirchhoff's integral $u_K(P)$ and its counterpart $\hat{u}_K(P)$ are reversed on opposite sides of the screen, so that they counterbalance each other and the discontinuities of the Rayleigh-Sommerfeld integrals $u_{RS}^{(p,s)}(P)$ are eliminated.

The general properties of the diffracted field defined by these quantities can be inferred from the above-

mentioned fact that the differences between the Rayleigh-Sommerfeld integrals are pronounced only in the immediate proximity of the screen and vanish in the mid zone. Thus the forward component $\hat{u}_K(\hat{P})$ in the lower Eq. (6i) disappears as the Fresnel limit is approached, and the reverse component $\hat{u}_K(P)$ in the upper equation (6j) is an evanescent wave which also vanishes in the Fresnel limit. For $|z| \gg \lambda$, the forward field is the same as in Fresnel's theory and the reverse field is a mirror image of the forward Fresnel pattern.

3.3 Implementation
3.3.1 General

The formulae of the preceding section are unsuited for numerical computations unless they can be reduced to single integrals that can be evaluated without simplifying assumptions which might degrade their accuracy. Fortunately, the rare instances in which this can be done include two cases of great practical importance: circular apertures and apertures bounded by straight edges when illuminated by normally incident light. The single integrals obtained in these two cases can readily be evaluated by numerical methods based on the algorithms described in Sects. 3.3.2 and 3.3.3, below. The use of these algorithms on a personal computer with standard spreadsheet software is straightforward and yields numerical results that were found to be everywhere finite except at distances $|z| < 0.01\lambda$ from the aperture plane.

The same algorithms can also used at distances $z \gg \lambda$, and this is recommended because the standard analytical solutions of Fresnel's integral (2) are notoriously difficult to use on a personal computer. For example, the algorithm defined by Eqs. (8a-c), below, is less tedious than implementing Lommel's analytical solution of the Fresnel diffraction pattern of circular apertures in terms of infinite series of Bessel functions [2]. Likewise, the use of Eqs. (11a,b) is easier than the computation of complex Fresnel integrals by Taylor series and polynomial approximations [14]. The algorithms (8a-c) and (11a,b) have the added advantage of being rigorously accurate, and it should also be noted that in most cases it is not necessary to compute the reverse field. On the other hand, the methods described in this paper can only be used for normally incident light. It should also be mentioned that, owing to the highly structured nature of the diffraction patterns, the use of any of these methods can be cumbersome when wide apertures are considered.

3.3.2 Circular Apertures
3.3.2.1 Algorithms

Let $ABCB'A'$ be the rim of a circular aperture of radius w which is illuminated by normally incident coherent light and is centered on the coordinate origin, as shown in Fig. 2. As the corresponding diffraction pattern must be rotationally symmetrical about the z-axis it will be sufficient to consider its variation in the xz-plane, and thus the point of observation is chosen as $P = (x, 0, z)$. The integrals (5a,b) may then be reduced to single integrals by defining the area elements dQ so that they are all concentric with the projection $Q_0 = (x, 0, 0)$ of P onto the aperture plane and coincide with the circles $QBQ_\xi B'$ shown in the figure, where $Q_\xi = (\xi, 0, 0)$ is the right-most point at which these circles intersect the x-axis. Under these conditions the phases kQP will be constant and equal to

$$\beta \equiv k\,QP = k\,Q_\xi P = k\sqrt{(\xi - x)^2 + z^2} = \sqrt{v^2 + (kz)^2},$$
$$v = k(\xi - x) \tag{7a}$$

everywhere on these area elements and the integration can be carried out over $\xi - x$ alone. As also indicated in the figure, these area elements are in general not fully contained in the aperture and must therefore be evaluated as

$$dQ = 2\pi\,d(\xi - x)(\xi - x)(1 - \chi/\pi)$$
$$= (2\pi/k^2)dv\,v\,(1 - \chi/\pi), \tag{7b}$$

where 2χ is the angle subtended by the obstructed arc $BQ_\xi B'$ and is given by

$$\cos\chi = \frac{w^2 - x^2 - (\xi - x)^2}{2x\,(\xi - x)}$$
$$= \frac{(kw)^2 - (kx)^2 - v^2}{2kxv}, \tag{7c}$$

or $\eta = 0$ or π, as appropriate, when the right-hand side of (7c) exceeds ± 1. Hence one finds, using Eqs. (5a,b) and (7a,b),

$$u_{RS}^{(p)}(x, z) = -\frac{ik^2}{2\pi}\int dQ\,\frac{\exp(i\beta)}{\beta}$$
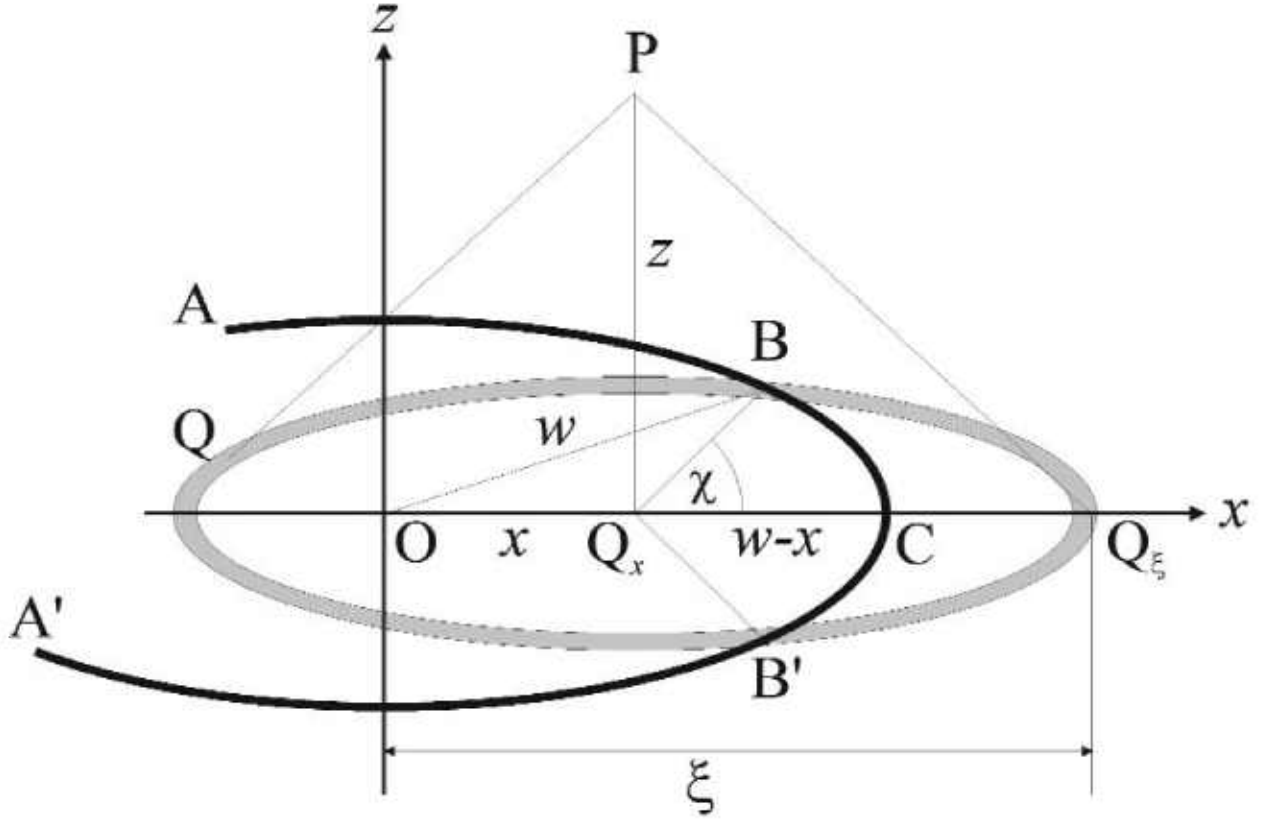$$= -i\int dv\,v\,(1 - \chi/\pi)\frac{\exp(i\beta)}{\beta}, \quad z > 0, \tag{7d}$$

Fig. 2. Notation used for circular apertures.

$$u_{RS}^{(s)}(x,z)=\frac{1}{ik}\frac{\partial u_{RS}^{(s)}(x,z)}{\partial z}$$

$$=-kz\int dv\, v\,(1-\chi/\pi)(i-\frac{1}{\beta})\frac{\exp(i\beta)}{\beta}, \quad (7e)$$

$z>0$,

so that the integrals defined in Eq. (6e) are now given by

$$u_K(x,z)=\int dv\, v\,(1-\chi/\pi)\,C(\beta), \quad z>0,$$

$$C(\beta)=[(\frac{1}{\beta}+\frac{kz}{\beta^2})\sin\beta+\frac{kz}{\beta^3}\cos\beta] \quad (7f)$$

$$-i[(\frac{1}{\beta}+\frac{kz}{\beta^2})\cos\beta-\frac{kz}{\beta^3}\sin\beta],$$

$$\hat{u}_K(x,z)=\int dv\, v\,(1-\chi/\pi)\,\hat{C}(\beta), \quad z>0,$$

$$\hat{C}(\beta)=[(\frac{1}{\beta}-\frac{kz}{\beta^2})\sin\beta-\frac{kz}{\beta^3}\cos\beta] \quad (7g)$$

$$-i[(\frac{1}{\beta}-\frac{kz}{\beta^2})\cos\beta+\frac{kz}{\beta^3}\sin\beta].$$

The limits of these integrals are $v=0$ to $k(w+x)$ when $x\le w$, and $v=k(x-w)$ to $k(x+w)$ when $x\ge w$. In the first of these ranges it is assumed that $\chi\equiv 0$ when $v\le k(x-w)$.

In order to evaluate these integrals numerically, divide the aperture radius w into N equal elements, and let $\xi-x=nw/N$, $x=mw/N$. Hence, $dv=kw/N$, $v=nkw/N$, and therefore

$$u_K(x,z)=u_K(m,z)=(\frac{kw}{N})^2\sum_n (n-\tfrac{1}{2})\,(1-\chi_{n,m})\,C\,(\beta_n),$$

$$z>0, \quad (8a)$$

$$\hat{u}_K(x,z)=\hat{u}_K(m,z)=(\frac{kw}{N})^2\sum_n (n-\tfrac{1}{2})\,(1-\chi_{n,m})\,\hat{C}\,(\beta_n),$$

$$z>0, \quad (8b)$$

where

$$\chi_{n,m}=0 \quad \text{if } n\le m, \qquad \chi_{n,m}=\cos^{-1}\frac{N^2-m^2-(n-\tfrac{1}{2})^2}{2m(n-\tfrac{1}{2})}$$

$$\text{if } n>m, \quad (8c)$$

$$\beta_n = \sqrt{\frac{(n-\frac{1}{2})^2 (kw)^2}{N^2} + (kz)^2} , \qquad (8d)$$

all quantities being evaluated at the mid-points of the summation elements. The ranges of summation in Eqs. (8a) are $1 \le n \le m + N$ if $n \le m$ and $m - N < n \le m + N$. The number N of summation elements used in these expressions must be large enough to ensure that the oscillations of $e^{i\beta}$ in Eqs. (7d,e) are accurately sampled. According to Eq. (7a) and the quarter-wave criterion this will be achieved if the pathlength difference ΔPQ between successive summation elements is less than $\lambda/4$, or $\Delta\beta \le \pi/2$. Differentiating Eq. (7a) with respect to v gives

$$\frac{kw}{N} = \Delta v = \frac{\beta\,\Delta\beta}{v} < \frac{\pi\beta}{2v} , \quad N > \frac{2v\,kw}{\pi\beta} , \qquad (8e)$$

which can now be used as follows to estimate the required value of N. In the immediate vicinity of the screen ($z \to 0$ we have $\beta \sim v$, so that $N > 2w/\lambda = 100$ when $w = 50\,\lambda$. In the Fresnel limit ($z \gg \lambda$) one finds $N > 2\,v_{max}w/\pi z$ where v_{max} is the largest value of v used in the computations, and therefore $N > 6w^2/\lambda z = 150$ when $w = 50\,\lambda$, $z = 100\,\lambda$ and the farthest point of observation is located one aperture halfwidth beyond the shadow boundary ($v_{max} = 1.5\,kw$). It should be noted that the corresponding values of N can be much larger for wider apertures.

3.3.2.2 Numerical Examples

1. Forward and Reverse Axial Irradiances. Equations (7f,g) can be solved in closed form for the special case of axial points of observation, where $x = 0$, $\chi = 0$ and

$\beta = k\sqrt{\xi^2 + z^2}$ and therefore

$$u_{RS}^{(p)}(0,z) = -ik^2 \int_0^{kw} d\xi\,\xi\,\frac{\exp(i\beta)}{\beta}$$

$$= -\int_{kz}^{kW} d\,t\exp(t) = \exp(i\,kz) - \exp(i\,kW) , \qquad (9a)$$

$$t = i\beta , \quad W = \sqrt{w^2 + z^2} .$$

Hence one finds, using Eqs. (7e) and (6e),

$$u_{RS}^{(s)}(0,z) = \exp(i\,kz) - \frac{z\exp(i\,kW)}{W} ,$$

$$u_K(0,z) = \exp(i\,kz) - \frac{1}{2}\exp(i\,kW)(1 - \frac{z}{W}) , \quad (9b)$$

$$\hat{u}_K(0,z) - \frac{1}{2}\exp(i\,kW)(1 + \frac{z}{W}) ,$$

which can now be substituted into Eqs. (6i,j) and then leads to the following expressions for the forward and reverse irradiances along the z-axis of Fig. 2,

$$E(0,z) = E_{geom}\left\{1 + \frac{1}{4}(1 - \frac{|z|}{W})^2 \right.$$

$$\left. - \cos[k(z-W)](1 - \frac{|z|}{W})\right\} , \qquad (9c)$$

$$\hat{E}(0,z) = \frac{1}{4}E_{geom}(1 - \frac{|z|}{W})^2 .$$

These results are valid for arbitrary values of z and are plotted in Fig. 3, where the upper curve represents the forward axial irradiance $E(0, z)$ computed from the first Eq. (9c) for $w = 5\,\lambda$ and the lower curve represents the corresponding reverse irradiance $\hat{E}(0, z)$. Both quantities are continuously differentiable on crossing the aperture plane, the forward irradiance (a) being equal to the geometrical irradiance E_{geom} in the negative Fresnel limit $-z \gg \lambda$, oscillating rapidly in the vicinity
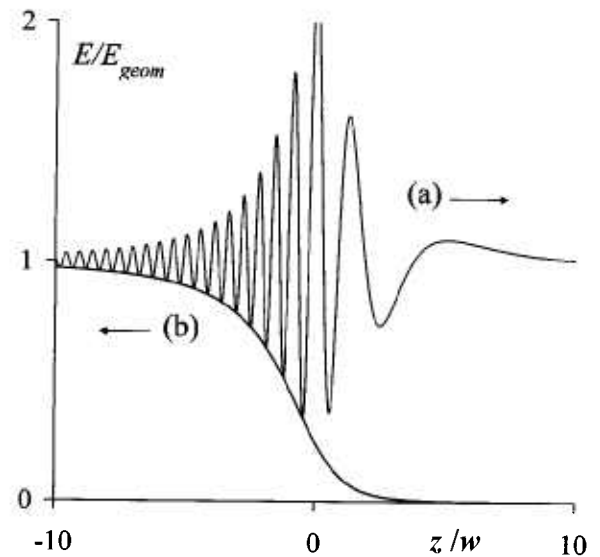

Fig. 3. Forward and reverse axial irradiances, (a) E_K and (b) $\hat{E}_K$, on opposite sides of a circular aperture of radius $w = 5\,\lambda$.

of the aperture, and tapering off beyond it. The reverse axial irradiance (b) is seemingly diverging from a virtual source point beyond $z = -10\lambda$, exhibits no oscillations, and its magnitude is equal to the lower envelope of the forward irradiance curve.

2. Near-Field Diffraction Patterns. The application of Eqs. (8a-c) for the computation of diffraction patterns is straightforward, and in this work standard spreadsheet software was used to obtain numerical results. As an example, Fig. 4 shows the near-field irradiance profiles (6i) on the opposite sides ($z = \pm \lambda$) of

the previously considered circular aperture of radius $w = 5\lambda$. The resemblance of the central portions of these curves is remarkable and is attributable to the basic premise adopted in Sec. 3.2, where St. Venant's hypothesis was replaced with the assumption that diffraction is a continuous field phenomenon that occurs on both sides of the screen. The corresponding reverse profiles defined by Eq. (6j) are shown in Fig. 5, illustrating the onset of a reverse flow of energy on the positive side of the screen ($z = \lambda$) as well as the fact that the reflected diffraction pattern on the negative side ($z = -\lambda$) is simply a mirror image of the transmitted pattern in Fig. 4.
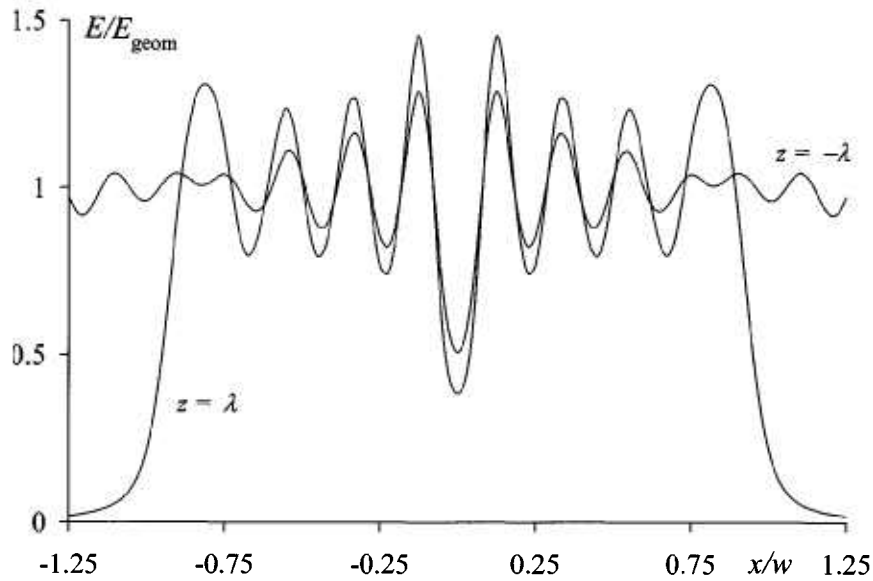

Fig. 4. Forward irradiance profiles E_K at distances $z = \pm \lambda$ from a circular aperture of radius $w = 5\lambda$.
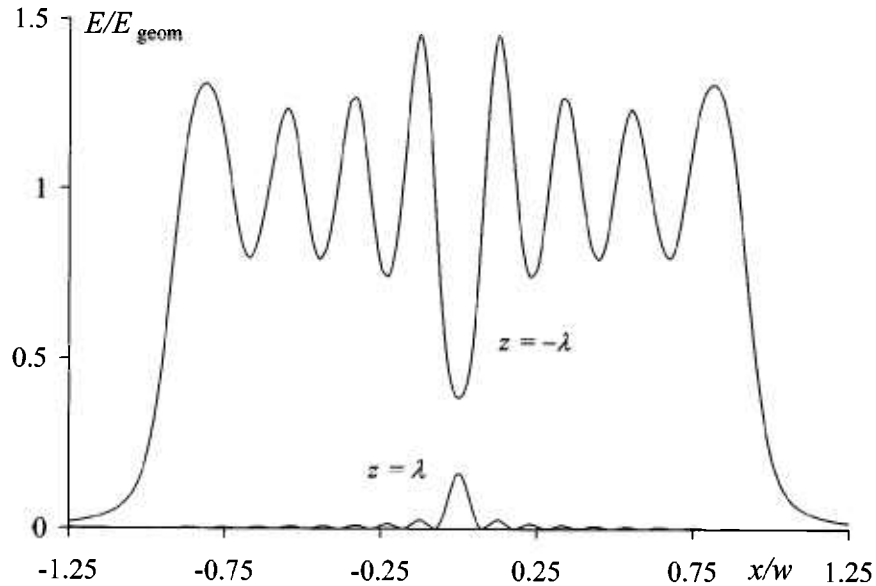

Fig. 5. Reverse irradiance profiles $\hat{E}_K$ at distances $z = \pm \lambda$ from a circular aperture of radius $w = 5\lambda$.

3. Aperture Field and Transmission Coefficients. For radiometric applications it is important to know the transmission coefficient of an aperture, defined as $\tau = \Phi_{total} / \Phi_{geom}$ where Φ_{total} is the total radiant flux transmitted into the half space $z > 0$ and Φ_{geom} is the geometrical flux incident upon it in the absence of diffraction. Thus, for a circular aperture as discussed in this Section,

$$\tau = \frac{1}{AE_{geom}} \int_{\mathscr{A}} dQ \; E(Q) = \frac{2}{w^2} \int_0^w dx \, x \, | u_K(x, 0)|^2$$

$$= \frac{2}{N^2} \sum_{m=1}^{N} (m - \tfrac{1}{2}) \, |u_K(m, 0)|^2 , \qquad (10a)$$

where $A = \pi w^2$ is the aperture area, $dQ = 2\pi x \, dx$ is the circular area element, and $u_K(m, z)$ is given by Eq. (8a). The practical use of Eq. (10a) is tedious because it requires consecutive numerical integrations and also poses computational problems arising from the singularities of $u_K(x, z)$ in the limit $z \to 0$. The most troublesome singularities, due to the terms in $1/\beta^2$ and $1/\beta^3$ in Eq. (7f), can be avoided altogether by invoking the second Eq. (5e) so that

$$u_K(m, 0) \approx \tfrac{1}{2}[1 + u_{RS}^{(p)}(m, z_{min})], \qquad (10b)$$

where $u_{RS}^{(p)}$ is singular in $1/\beta$, only, and z_{min} is very small. Trial computations indicated that values of z_{min} as small as $0.0003 \, \lambda$ could be used without difficulty and that the limiting value of τ defined by Eq. (10a) was reached at the 0.1 % level for $z_{min} < 0.0003 \, \lambda$. Accordingly, the numerical result presented in the following were computed for $z_{min} = 0.001 \, \lambda$. As expected, the aperture irradiance distributions $|u_K(x, 0)|^2$ obtained from Eq. (10b) were similar to an average of the two curves in Fig. 5. They were everywhere finite and continuous and bore no similarity to the aperture distributions presumed by Rayleigh and Bouwkamp [6,10], Levine and Schwinger [11] or Wolf and Marchand [15].

Figure 6 shows the dependence of the transmission coefficient (10a) on aperture size for the range $0 < kw < 3\pi$, and here it is seen that τ exhibits a damped oscillatory behavior and quickly approaches the limit, $\tau \to 1$ as $kw \to \infty$. It was estimated that this limit is reached within less than 1 % when $w = 5 \, \lambda$.
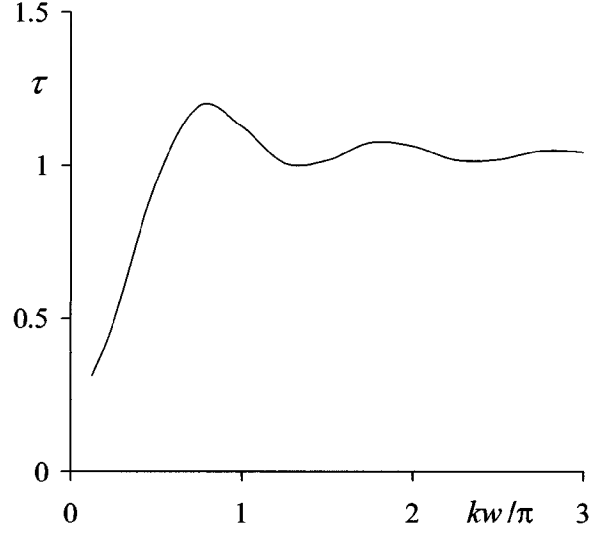

Fig. 6. Transmission coefficient τ of circular apertures vs aperture radius w.

3.3.3 Apertures Bounded by Straight Edges
3.3.3.1 Algorithms

Consider a plane aperture of width $(l + r)$, bounded by parallel straight edges as indicated in Fig. 7. The corresponding diffraction pattern will consist of straight bands which are parallel to the edges, and thus it will again be sufficient to compute its variation along the x-axis. For a given point of observation $P = (x, 0, z)$ and arbitrary aperture points $Q = (\xi, \eta, 0)$, Eqs. (5a,b) can now be expressed as follows [12],

$$u_{RS}^{(p)}(x, z) = -\frac{ik}{2\pi} \int_{-l}^{r} d\xi \int_{-\infty}^{\infty} \frac{d\eta \, \exp[ik\sqrt{(\xi - x)^2 + \eta^2 + z^2}]}{\sqrt{(\xi - x)^2 + \eta^2 + z^2}}$$

$$= \frac{1}{2} \int_{-k(l+x)}^{k(r-x)} dv \, H_0^{(1)}(\beta) ,$$

$$u_{RS}^{(s)}(x, z) = \frac{ikz}{2} \int_{-k(l+x)}^{k(r-x)} \frac{dv \, H_1^{(1)}(\beta)}{\beta} , \qquad (11a)$$

where $z > 0$, $H_0^{(n)} = J_n + iY_n$ denotes a Hankel function, J_n and Y_n are Bessel functions, and v as well as β are the same as in Eq. (7a). It follows at once that the forward and reverse wave functions defined in Sec. 3.2 are given by
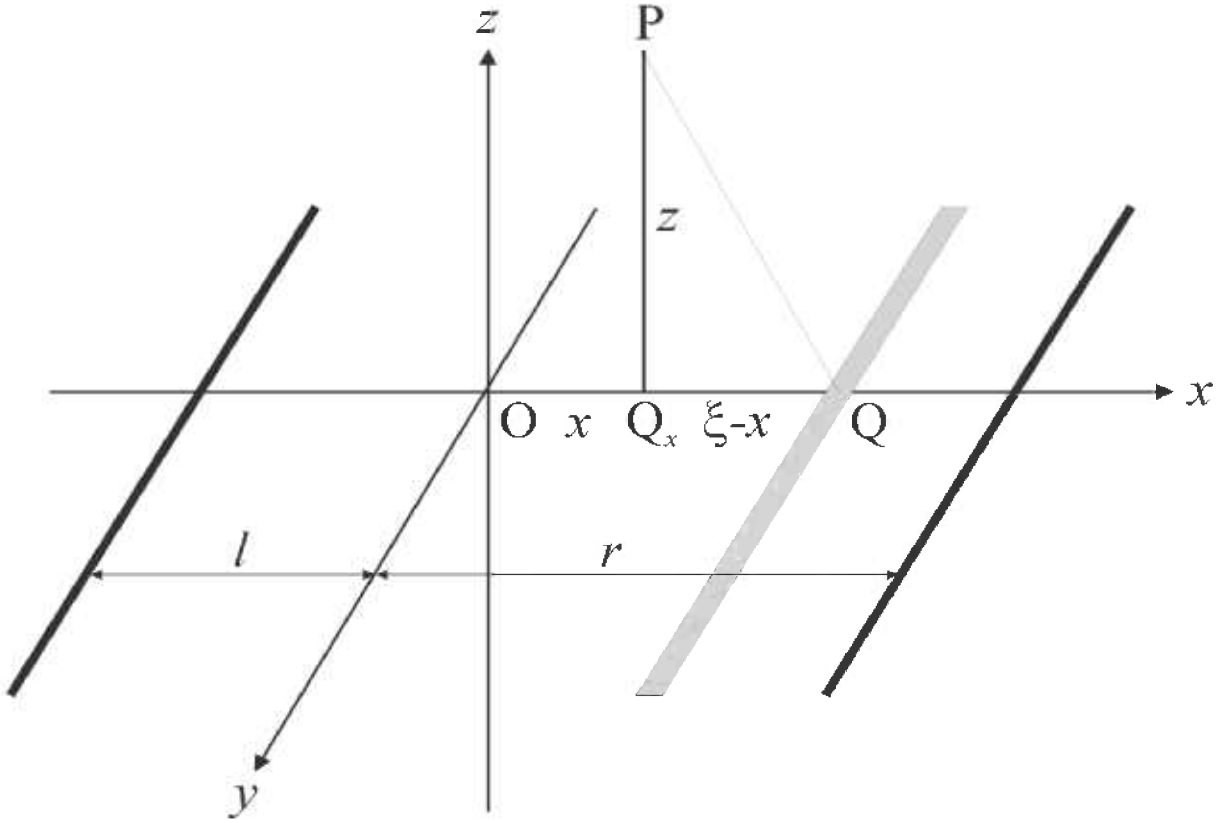
Fig. 7. Notation used for slits.

$$u_K(x,z) = \frac{1}{2}[u_{RS}^{(p)}(x,z) + u_{RS}^{(s)}(x,z)]$$

$$= \int_{-k(l+x)}^{k(r-x)} dv \, S(\beta), \quad z > 0,$$

$$S(\beta) = \frac{1}{4}[J_0(\beta) - \frac{kz}{\beta}Y_1(\beta]$$

$$+ \frac{i}{4}[Y_0(\beta) + \frac{kz}{\beta}J_1(\beta], \quad (11b)$$

$$\hat{u}_K(x,z) = \frac{1}{2}[u_{RS}^{(p)}(x,z) - u_{RS}^{(s)}(x,z)]$$

$$= \int_{-k(l+x)}^{k(r-x)} dv \, \hat{S}(\beta), \quad z > 0,$$

$$\hat{S}(\beta) = \frac{1}{4}[J_0(\beta) + \frac{kz}{\beta}Y_1(\beta] \quad (11c)$$

$$+ \frac{i}{4}[Y_0(\beta) - \frac{kz}{\beta}J_1(\beta].$$

To evaluate these expressions by numerical integration define, in analogy to the definitions that precede Eqs. (8a), $l+r = 2w$, $\Delta v = kw/N$, $L = 2kl/\Delta v$, $R = 2kr/\Delta v$, $v = n\Delta v$, $kx = m\Delta v$. Therefore,

$$u_K(x,z) = u_K(m,z) = \frac{kw}{N} \sum_{n=-(L+m)}^{R-m} S(\beta_n), \quad z > 0, \quad (12a)$$

$$\hat{u}_K(x,z) = \hat{u}_K(m,z) = \frac{kw}{N} \sum_{n=-(L+m)}^{R-m} \hat{S}(\beta_n), \quad z > 0, \quad (12b)$$

where, as before, $\beta_n = \sqrt{[(n-\frac{1}{2})^2(kw)^2/N^2] + (kz)^2}$ and the choice of N is governed by the same considerations as in Sec. 3.3.2.1. As β_n is independent of m it follows at once that, if $u_K(m,z)$ and $\hat{u}_K(m,z)$ are known and m is replaced by $m \pm 1$, the new values will be

$$u_K(m \pm 1, z) = u_K(m,z) \mp \frac{kw}{N}[S(\beta_{-L+m\pm 1}) - S(\beta_{R-m\pm 1})],$$

$$z > 0, \quad (12c)$$

$$\hat{u}_K(m\pm 1, z) = \hat{u}_K(m, z) \mp \frac{kw}{N}[\hat{S}(\beta_{-L+m\pm 1}) - \hat{S}(\beta_{R-m\pm 1})],$$

$$z > 0, \tag{12d}$$

which illustrates in a very instructive manner how the diffraction pattern changes when the point of observation is moved so that new portions of the wavefront are covered and uncovered by the aperture edges. The recursion formulae (12c,d) are convenient for practical applications as they allow the computation of successive values without performing the summations of Eqs. (12a,b) for every point of observation.

The results obtained for the diffraction profiles of slits are similar to those presented in Sec. 3.3.2.2 for circular apertures. They were included in Ref. [12] and are omitted here.

3.4 Half-Plane Results and Comparison With Sommerfeld's Theory

The aperture depicted in Fig. 7 is reduced to an infinitesimally thin half plane defined by $x > 0$, $z = 0$, by letting $L = -\infty$, $R = 0$. Accordingly, Eqs. (12a-d) are now replaced by

$$u_K(x, z) = \int_{-\infty}^{-kx} dv\, S(\beta) = u_K(m, z) = \Delta v \sum_{n=-\infty}^{-m} S(\beta_n),$$

$$u_K(m\pm 1, z) = u_K(m, z) \mp \Delta v\, S(\beta_{m\pm 1}), \quad z > 0, \tag{13a}$$

$$\hat{u}_K(x, z) = \int_{-\infty}^{-kx} dv\, \hat{S}(\beta) = \hat{u}_K(m, z) = \Delta v \sum_{n=-\infty}^{-m} \hat{S}(\beta_n),$$

$$\hat{u}_K(m\pm 1, z) = \hat{u}_K(m, z) \mp \Delta v\, \hat{S}(\beta_{m\pm 1}), \quad z > 0, \tag{13b}$$

where it should be noted that the last terms on the right-hand sides of Eqs. (12c,d) are now absent because there is no right aperture edge. Here, as above, $kx = m\Delta v$, $k(\xi - x) = n\Delta v$, $S(\beta_n)$ and $\hat{S}(\beta_n)$, being the same as in (11b,c), and $\beta_n = \sqrt{(n-\frac{1}{2})^2 \Delta v^2 + (kz)^2}$. The summation elements Δv must again satisfy the quarter-wave criterion, so that the phase difference between successive summation elements, $\Delta\beta = \beta_{n+1} - \beta_n$, must not exceed $\pi/2$. This phase difference has a maximum value, $(\Delta\beta)_{max} = \Delta v$ when $z = 0$, and hence it follows that the quarter-wave condition will always be satisfied when $\Delta v < \pi/2$. As the choice of Δv also determines the step size of the recursions (13a,b) the value chosen in this work was $\Delta v = \pi/5$, yielding equidistant values

of $u_K(x, z)$ and $\hat{u}_K(x, z)$ spaced by $\Delta x = \pi/5k = 0.1\lambda$. The starting values, $u_K(0, z) = 0.5$ and $\hat{u}_K(0, z) = 0$, were obtained by performing the summations in Eqs. (13a,b) for $m = 0$.

The corresponding expressions according to Sommerfeld's rigorous theory [7, 9, 12] for the border case of a normally incident geometric field are

$$u_S^{(p,s)}(x, z) = u_S(x, z) \pm \hat{u}_S(x, z), \quad z \neq 0, \tag{14a}$$

$$u_S(x, z) = \exp(ikz)\, V(\rho), \quad \rho = \sqrt{2/\lambda}\,[-\operatorname{sign}(z)\sqrt{r+x} + \sqrt{r-x}], \quad z \neq 0, \tag{14b}$$

$$\hat{u}_S(x, z) = \exp(-ikz)\, V(\hat{\rho}), \quad \hat{\rho} = \sqrt{2/\lambda}\,[-\operatorname{sign}(z)\sqrt{r+x} - \sqrt{r-x}], \quad z \neq 0, \tag{14c}$$

$$V(\rho) = \frac{\exp(-i\pi/4)}{\sqrt{2}} \int_{-\infty}^{\rho} d\tau \exp(i\pi\tau^2/4) = \frac{1}{2}[1 + C(\rho) + S(\rho)] - \frac{i}{2}[C(\rho) - S(\rho)], \tag{14d}$$

where $u_S^{(p,s)}(x, z)$ satisfy the wave equation and the respective boundary conditions $\partial u_S^{(p)}/\partial z = 0$ and $u_S^{(s)} = 0$, u_S and $\hat{u}_S$ are the forward and reverse wave functions, $V(\rho)$ is the complex Fresnel integral the form used by Sommerfeld,[6] while $C(\rho)$ and $S(\rho)$ are the standard Fresnel cosine and sine integrals.

During the course of this work, the above sets of equations were used for numerous computations intended to quantify the residual differences between them. The results obtained showed consistently that these differences were insignificant even in the immediate vicinity of the diffracting edge. As an example, Fig. 8 shows the respective forward irradiance distributions, $|u_K|^2$ and $|u_S|^2$, at the distances $z = \pm\lambda$ from the half plane. The similarity of the curves is unmistakable, and further computations showed that the same quantities differ by less than ± 0.01 for $z = \pm 10\lambda$. This demonstrates the rigor of the equations derived in Sec. 3.2 for the specific case of half-plane diffraction, and it may be inferred that the corresponding expressions for circular apertures and slits are equally as rigorous.

---

[6] Different notations have been used elsewhere. For example, in Ref. [17] therse relationships were expressed of the function

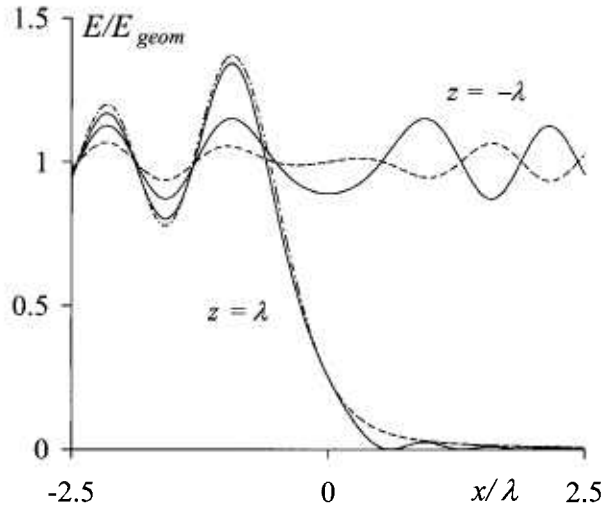$$F(-\sqrt{\pi/2}\,\rho) = \exp(i\pi/4)\sqrt{\pi}\, V(\rho).$$

Fig. 8. Forward irradiance profiles E_K at distances $z = \pm\lambda$ from a half plane according to Eqs. (13a) (-----) and (14b) (——).

4. A Note on Electromagnetic Diffraction Theories

The thoughts presented in this paper arose from a realization that, at some time in the mid 1900's, the theory of diffraction had reached an impasse. Although it should have been obvious that the inability to account for polarization phenomena is an intrinsic feature of the scalar theory of light, and although there was no experimental evidence that this is a practical problem in the context of Fresnel's theory,[7] attempts were made to modify the theory so that, in a manner of speaking, polarization is introduced through the back door and the scalar theory assumes a pseudo-vectorial character. These modifications were justified as follows [e.g., 10, 16].

For infinite plane screens and plane-parallel incident light, the Helmholtz integral theorem [5] has two mutually independent solutions, $u^{(p)}$ and $u^{(s)}$, which include but are not limited to the Rayleigh-Sommerfeld and Sommerfeld solutions $u_{RS}^{(p,s)}$ and $u_S^{(p,s)}$ discussed in

Sec. 3 and obey the same boundary conditions as those pertaining to the reflection of polarized light at a perfectly conducting metallic reflector. Hence it is surmised, without further proof, that these solutions can be regarded as the components of mutually independent electromagnetic fields given by $\mathbf{H}^{(p)} = \sqrt{\varepsilon/\mu}(0, u^{(p)}, 0)$, $i k \mathbf{E}^{(p)} = -\sqrt{\mu/\varepsilon}\, \mathrm{curl}\,\mathbf{H}^{(p)}$ and $\mathbf{E}^{(s)} = \sqrt{\mu/\varepsilon}(0, u^{(s)}, 0)$, $i k \mathbf{H}^{(s)} = \sqrt{\varepsilon/\mu}\,\mathrm{curl}\,\mathbf{E}^{(s)}$, where SI units are used and $\sqrt{\mu/\varepsilon}$ is the wave impedance of free space.

It is easy to show that these assumptions are fallacious and can lead to contradictory or dubious results. For example:

According to the above, unpolarized light incident on a diffracting screen is transmitted as partially polarized light having a polarization ratio given by $\Pi = |u^{(p)}|^2 / |u^{(s)}|^2$ and ordinarily this ratio would not be unity in the near zone. However, in the mid zone $u^{(p)}$ and $u^{(s)}$ are expected to be same so that $\Pi = 1$. This is absurd, because the polarization of light cannot change while propagating in free space.

The curl operators in Maxwell's equations are known to introduce singularities into the "electromagnetic" theory, even when the underlying wave functions $u^{(p)}$ and $u^{(s)}$ are everywhere finite and continuous. These singularities have been interpreted as evanescent edge waves emitted by the diffracting edge, as originally presumed by Thomas Young but soon refuted by Fresnel.[8] The notion of edge waves has persisted as a means to explain singularities encountered in diffraction calculations [18], whereas a thorough analysis might have revealed that these singularities are artifacts of mathematical errors or illogical assumptions.

The reason for inconsistencies of this kind is that the conditions $u^{(p)} = 0$ and $\partial u^{(s)}/\partial z = 0$ are necessary and sufficient to satisfy Helmholtz' theorem, but insufficient to transform the scalar theory of light into a viable tool for explaining electromagnetic phenomena.

---

[7] This may be due to the paraxial approximations made in Fresnel's theory. The existence of polarization effects at large angles is, of course, evidenced by the properties of diffraction gratings.

[8] It appears that Young did not publish this hypothesis and that Fresnel was the first to mention it.

5.   References

[1] A. J. Fresnel, Mémoire sur la Diffraction de la Lumière (Paris 1819). Transl.: H. Crew et al, The Wave Theory of Light and Spectra, Arno Press, pp. 79-144 (1981).

[2] K. D. Mielenz, Algorithms for Fresnel Diffraction at Rectangular and Circular Apertures, J. Res. Natl. Inst. Stand. Technol. **103**, 497-509 (1998).

[3] K. D. Mielenz, Wolf Shifts and their Physical Interpretation under Laboratory Conditions, J. Res. Natl. Inst. Stand. Technol. **98**, 231-240, Sec.5.1. (1993).

[4] J. G. Suárez-Romero, E. Tepechín-Rodríguez and K. D. Mielenz, Cross Spectral Density Propagated through a Circular Aperture, Metrologia **38**, 379-384 (2001).

[5] H. L. F. v. Helmholtz, Theorie der Luftschwingungen in Röhren mit offenen Enden,: J. Mathematik **57**, 1-72 (1859).

[6] Lord Rayleigh, On the Passage of Waves through Apertures in Plane Screens, Phil. Mag. **43**, 259-272 (1897); On the Passage of Waves through Fine Slits in Thin Opaque Screens, Proc. Roy. Soc. (A) **89**, 194-219 (1913).

[7] A. Sommerfeld, Optik, (Dieterich, 1950). Transl.: Theory of Optics, Longmans (1964).

[8] G. Kirchhoff, Zur Theorie der Lichtstrahlen, Ann. Physik **18**, 663-695 (1883).

[9] A. Sommerfeld, Mathematische Theorie der Diffraction, Math. Ann. **47**, 317-374 (1896).

[10] C. J. Bouwkamp, Diffraction Theory, Rep. Progr. Phys. (London) **17**, 35-100 (1953).

[11] H. Levine and J. Schwinger, On the Theory of Diffraction by an Aperture in an Infinite Plane screen, Phys. Rev. **74**, 958-974 (1948), ibid **75**, 1423-1432 (1949).

[12] K. D. Mielenz, Optical Diffraction in Close Proximity to Plane Apertures, J. Res. Natl. Inst. Stand. Technol. **107**, 355-362 (2002), ibid **108**, 57-68 (2003), ibid **109**, 457-464 (2004), ibid **111**, 1-8 (2006)

[13] J. Meixner, Die Kantenbedingung in der Theorie der Beugung elektromagnetischer Wellen an vollkommen leitenden ebenen Schirmen, Ann. Physik (6) **6**, 2-9 (1949).

[14] K. D. Mielenz, Computation of Fresnel Integrals II, J. Res. Natl. Inst. Stand. Technol. **105**, 589-590 (2000).

[15] E. Wolf and E. W. Marchand, Comparison of the Kirchhoff and the Rayleigh-Sommerfeld Theories of Diffraction at an Aperture, J. Opt. Soc. Am. **54**, 587-594 (1964).

[16] M. Born and E. Wolf, Principles of Optics (Pergamon, 1970).

[17] E. W. Marchand and E. Wolf, Boundary-Diffraction Wave in the Domain of the Rayleigh-Kirchhoff Diffraction Theory, J. Opt. Soc. Am. **52**, 761-767 (1962).

About the author: Klaus D. Mielenz is a physicist, Fellow of the Optical Society of America, and retired Chief of the Radiometric Physics Division of the National Institute of Standards and Technology. NIST is an agency of the U.S. Department of Commerce.